04-11-05

PTO/SB/21 (04-04)

OIPE JC10 APR 07 2005 PATENT & TRADEMARK OFFICE

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

Application Number	10/783,753
Filing Date	February 20, 2004
First Named Inventor	DAMIEN LE MOAL
Art Unit	2186
Examiner Name	Unassigned
Attorney Docket Number	16869P-105000US

Total Number of Pages in This Submission

ENCLOSURES *(Check all that apply)*

- ☒ Fee Transmittal Form (in duplicate)
- ☐ Fee Attached
- ☐ Preliminary Amendment
- ☐ After Final
- ☐ Affidavits/declaration(s)
- ☐ Extension of Time Request
- ☐ Express Abandonment Request
- ☐ Information Disclosure Statement
- ☐ Certified Copy of Priority Document(s)
- ☐ Response to Missing Parts/ Incomplete Application
- ☐ Response to Missing Parts under 37 CFR 1.52 or 1.53
- ☐ Drawing(s)
- ☐ Licensing-related Papers
- ☒ Petition to Make Special (10 pages)
- ☐ Petition to Convert to a Provisional Application
- ☐ Power of Attorney, Revocation Change of Correspondence Address
- ☐ Terminal Disclaimer
- ☐ Request for Refund
- ☐ CD, Number of CD(s) _____
- ☐ After Allowance Communication to Technology Center (TC)
- ☐ Appeal Communication to Board of Appeals and Interferences
- ☐ Appeal Communication to TC **(Appeal Notice, Brief, Reply Brief)**
- ☐ Proprietary Information
- ☐ Status Letter
- ☒ Other Enclosure(s) (please identify below):

Return Postcard

Nine (9) cited references

Remarks: The Commissioner is authorized to charge any additional fees to Deposit Account 20-1430.

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm or Individual name	Townsend and Townsend and Crew LLP Chun-Pok Leung Reg. No. 41,405
Signature	
Date	April 7, 2005

CERTIFICATE OF TRANSMISSION/MAILING

Express Mail Label: **EV 530888583 US**

I hereby certify that this correspondence is being deposited with the United States Postal Service with "Express Mail Post Office to Address" service under 37 CFR 1.10 on this date September 13, 2004 and is addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below.

Typed or printed name	Joy Salvador		
Signature		Date	April 7, 2005

60463575 v1

BEST AVAILABLE COPY

OIPE JC76 APR 07 2005 PATENT & TRADEMARK OFFICE

FEE TRANSMITTAL for FY 2004

Effective 10/01/2003. Patent fees are subject to annual revision.

☐ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT | ($) **130.00**

Complete if Known	
Application Number	10/783,753
Filing Date	February 20, 2004
First Named Inventor	DAMIEN LE MOAL
Examiner Name	Unassigned
Art Unit	2186
Attorney Docket No.	16869P-105000US

METHOD OF PAYMENT *(check all that apply)*

☐ Check ☐ Credit Card ☐ Money Order ☐ Other ☐ None

☒ Deposit Account:

Deposit Account Number: 20-1430

Deposit Account Name: Townsend and Townsend and Crew LLP

The Director is authorized to: *(check all that apply)*

☒ Charge fee(s) indicated below ☒ Credit any overpayments

☒ Charge any additional fee(s) or any underpayment of fee(s)

☐ Charge fee(s) indicated below, **except for the filing fee** to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1001	770	2001	385	Utility filing fee	
1002	340	2002	170	Design filing fee	
1003	530	2003	265	Plant filing fee	
1004	770	2004	385	Reissue filing fee	
1005	160	2005	80	Provisional filing fee	

SUBTOTAL (1) ($)0.00

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

			Extra Claims		Fee from below		Fee Paid
Total Claims		-** =		X		=	
Independent Claims		-** =		X		=	
Multiple Dependent				X		=	

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
1202	18	2202	9	Claims in excess of 20
1201	86	2201	43	Independent claims in excess of 3
1203	290	2203	145	Multiple dependent claim, if not paid
1204	86	2204	43	** Reissue independent claims over original patent
1205	18	2205	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($)0.00

**or number previously paid, if greater; For Reissues, see above*

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1051	130	2051	65	Surcharge - late filing fee or oath	
1052	50	2052	25	Surcharge - late provisional filing fee or cover sheet.	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	110	2251	55	Extension for reply within first month	
1252	420	2252	210	Extension for reply within second month	
1253	950	2253	475	Extension for reply within third month	
1254	1,480	2254	740	Extension for reply within fourth month	
1255	2,010	2255	1,005	Extension for reply within fifth month	
1401	330	2401	165	Notice of Appeal	
1402	330	2402	165	Filing a brief in support of an appeal	
1403	290	2403	145	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	110	2452	55	Petition to revive – unavoidable	
1453	1,330	2453	665	Petition to revive – unintentional	
1501	1,330	2501	665	Utility issue fee (or reissue)	
1502	480	2502	240	Design issue fee	
1503	640	2503	320	Plant issue fee	
1460	130	1460	130	Petitions to the Commissioner	130
1807	50	1807	50	Petitions related to provisional applications	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	770	2809	385	Filing a submission after final rejection (37 CFR § 1.129(a))	
1810	770	2810	385	For each additional invention to be examined (37 CFR § 1.129(b))	
1801	770	2801	385	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) _____

*Reduced by Basic Filing Fee Paid **SUBTOTAL (3)** ($)130.00

SUBMITTED BY Complete *(if applicable)*

Name (Print/Type)	Chun-Pok Leung	*Registration No. (Attorney/Agent)*	41,405	*Telephone*	650-326-2400
Signature				*Date*	April 7, 2005

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

OIPE JC10 APR 07 2005 PATENT & TRADEMARK OFFICE

PATENT
Attorney Docket No.: 16869P-105000US
Client Ref. No.: 340301089US01

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

DAMIEN LE MOAL et al.

Application No.: 10/783,753

Filed: February 20, 2004

For: STORAGE DEVICE ADAPTER EQUIPPED WITH INTEGRATED CACHE

Customer No.: 20350

Examiner: Unassigned

Technology Center/Art Unit: 2186

Confirmation No.: 7385

PETITION TO MAKE SPECIAL FOR NEW APPLICATION UNDER M.P.E.P. § 708.02, VIII & 37 C.F.R. § 1.102(d)

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a petition to make special the above-identified application under MPEP § 708.02, VIII & 37 C.F.R. § 1.102(d). The application has not received any examination by an Examiner.

(a) The Commissioner is authorized to charge the petition fee of $130 under 37 C.F.R. § 1.17(i) and any other fees associated with this paper to Deposit Account 20-1430.

04/13/2005 WABDELR1 00000067 201430 10783753
01 FC:1464 130.00 DA

(b) All the claims are believed to be directed to a single invention. If the Office determines that all the claims presented are not obviously directed to a single invention, then Applicants will make an election without traverse as a prerequisite to the grant of special status.

(c) Pre-examination searches were made of U.S. issued patents, including a classification search, a computer database search, and a keyword search. The searches were performed on or around March 15, 2005, and were conducted by a professional search firm, Kramer & Amado, P.C. The classification search covered Class 711 (subclass 112) for the U.S. and foreign subclasses identified above. The computer database search was conducted on the USPTO systems EAST and WEST. The keyword search was conducted in Class 711 (subclasses 113, 114, 137, 141, 147, 152, 167, and 202). The inventors further provided three references considered most closely related to the subject matter of the present application (see references #7-9 below), which were cited in the Information Disclosure Statement filed on February 20, 2004.

(d) The following references, copies of which are attached herewith, are deemed most closely related to the subject matter encompassed by the claims:

(1) U.S. Patent No. 6,101,588;

(2) U.S. Patent No. 6,237,046 B1;

(3) U.S. Patent No. 6,275,897 B1;

(4) U.S. Patent No. 6,606,715 B1;

(5) U.S. Patent Publication No. 2003/0149843 A1;

(6) U.S. Patent Publication No. 2004/0148473 A1;

(7) U.S. Patent No. 6,463,509 B1;

(8) International Patent Publication No. WO 03/017598; and

(9) Japanese Patent Publication No. JP 2001-051901.

(e) Set forth below is a detailed discussion of references which points out with particularity how the claimed subject matter is distinguishable over the references.

A. Claimed Embodiments of the Present Invention

The claimed embodiments relate to the retrieval and caching of data from a storage area network and in particular to a device adapter that enables the caching of data accessed from a storage without the support or alteration of an operating system (OS).

Independent claim 1 recites a device adapter that interconnects a host computer and a main storage, the device adapter including a cache storage interface device for accessing a cache storage, a main storage interface device for accessing the main storage via a network, and a cache controller that processes a data output request made by the host computer. The cache controller includes a processing unit that caches, in the cache storage, data retrieved from the main storage; a processing unit that manages use and allocation of storage regions of the cache storage; a processing unit that manages load and performance information relating to the main storage and the cache storage; a processing unit that determines whether or not the requested data is cached; and a processing unit that determines, when the requested data is cached, which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved.

With reference to the specification, FIG. 1 illustrates a device adapter 200, which is connected to common host computer 100. See paragraph [0029]. Device adapter 200 includes a cache controller 220 that includes a storage interface monitor 221 that monitors main storage interface device 240. See paragraph [0031]. Cache controller manages storage allocation of local cache storage 400. Id. Each element of device adapter 200 is realized by hardware or a processor executing a program. See paragraph [0032]. FIG. 2 illustrates flow of a read request wherein after a cache hit, a determination is made whether the cache storage interface better than a main storage. See step s103. With reference to FIG. 1, cache controller 220 determines whether cache storage interface device 250 or main storage interface device 240 is the best interface in accordance with statistical information and the load monitored by storage interface monitor 221. See paragraph [0039]. Storage monitor 221 then selects the main storage interface device 240 when a data transfer rate of cache storage 400 approaches a maximum. See paragraph [0041]. FIG. 3 illustrates an embodiment with a fiber channel interface device 240 communicating with storage area

network (SAN) 300, which is coupled to a RAID 500. See paragraph [0044]. SCSI interface device 250 then manages SCSI disk 400. See paragraph [0045]. In this embodiment, storage interface monitor 221 monitors data transfer rates, manages statistical data and performance of the storage devices. See paragraph [0046]. FIG. 4 illustrates a cache directory that monitors a final use time associated with each storage device. See paragraph [0048].

One of the benefits that may be derived is that data retrieved from storage unit connected through a network to a storage device can be effectively and flexibly cached in a device connected to a host computer, and the flexible sizing of cache storage, an efficient caching process that does not place a load on the system, and dynamic control of a data retrieval destination to reduce the processing time of an input/output request made by the host computer, become possible.

B. Discussion of the References

1. U.S. Patent No. 6,101,588

U.S. Patent No. 6,101,588 to Farley relates to a mass storage subsystem including a plurality of devices for use in a digital data processing system. FIG. 1 illustrates computer system 10 connected to a digital data storage subsystem 12. See col. 3, lines 9-23. The digital data storage system 12 comprises one or more data stores 20(1) and one or more host adapters (24(n). See col. 3, lines 24-35. Each of the host adapters 24(n) and each of the device controllers 21(m) include a cache manager 25(n) and 23(m), respectively, to access to the cache memory 31, cache index directory 32 and cache manager memory 33. See col. 4, line 23 to col. 5, line 8.

As understood, the cache controllers of Farley do not manage load and performance information relating to a main storage and a cache storage. Thus, Farley does not determine which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

2. U.S. Patent No. 6,237,046 B1

U.S. Patent No. 6,237,046 to Ohmura et al. relates to an input/output control apparatus, which has a channel adapter module coupled to a channel unit, a device adapter module coupled to a device, and a cache control module for managing a cache memory on

the basis of a hash table. FIG. 2 illustrates disk controller as an input/output control apparatus including channel adapter modules 16-1 and 16-2. See col. 11, lines 1-6. A cache function engine module 26 responds with a mishit in response to an inquiry of a cache status from a channel adapter module (16-1, 16-2). See col. 11, lines 6-16. The cache function engine module 26 includes a hash table 58 for executing input/output controls. Id. A storage area of the track data is then newly allocated onto the cache memory and information is stored in an allocation information region only for use by the channel adapter module. See col. 10, line 65 to col. 11, line 54.

As understood, the hash table of Ohmura et al. does not manage load and performance information relating to a main storage and a cache storage. Thus, Ohmura et al. does not determine which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

3. U.S. Patent No. 6,275,897 B1

U.S. Patent No. 6,275,897 to Bachmat relates to remote cache utilization for mirrored mass storage subsystem. FIG. 1 illustrates computer system 10 including a plurality of host computers 11(1) through 11(N). See col. 3, lines 46-51. A common memory subsystem 30 operates to cache information from the data stores 20(m) for access by the host computers 11(n) through the host adapters 24(n). See col. 4, lines 47-53. The common memory subsystem 30 includes a cache memory 31, a cache index directory 32 and a cache manager 33. See col. 7, line 30 to col. 8, line 41.

As understood, the cache manager of Bachmat does not manage load and performance information relating to a main storage and a cache storage. Thus, Bachmat does not determine which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

4. U.S. Patent No. 6,606,715 B1

U.S. Patent No. 6,606,715 to Kikuchi relates to a device control apparatus and control method for performing a cache writing operation. FIGS. 3A and 3B are block diagrams of a RAID controller 18-1 including cache control unit 40. See col. 6, line 60 to col. 7, line 9. FIG. 6 illustrates a protection management table 45 stored in cache

management area 48 of FIGS. 3A and 3B for storing protection data information. See col. 9, lines 40-50. If data cannot be written, an error message is sent to MPU 32. See col. 10, lines 29-32.

As understood, the cache control unit of Kikuchi manages stored protection data. Thus, Kikuchi does not determine which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

5. U.S. Patent Publication No. 2003/0149843 A1

U.S. Pub. No. 2003/0149843 to Jarvis et al. relates to a cache management system with multiple cache lists employing roving removal and priority-based addition of cache entries. FIG. 1 illustrates system 100 including a subsystem 104 (including an interface 106, storage/cache manager 108, cache 110, and metadata 112). See paragraph [0022]. A host 102 (which may be an application program) initiates a request for storage 116. Id. The subsystem 104 includes an interface 106 that cooperates with storage/cache manager 108. See paragraph [0024]. Cache 110 is a storage for use in caching more frequently or recently used data. Id. Table 1 illustrates an exemplary cache list including location of data in cache and time of most recent use. See paragraph [0031].

As understood, the cache list of Jarvis et al. does not manage load and performance information relating to a main storage and a cache storage. Thus, Jarvis et al. does not determine which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

6. U.S. Patent Publication No. 2004/0148473 A1

U.S. Pub. No. 2004/0148473 to Hughes et al. relates to a data processing system having a memory hierarchy including a cache and a lower-level memory system. FIG. 7 illustrates a portion 300 of central processing unit 122, including cache 124. See paragraph [0041]. Read requests are processed through a multiplexer 322 and if data is present in the cache, it results in a probe hit. Id. Cache 124 also receives read response data as a result of reads to a lower level memory system and stores such data in a read response

data buffer 318. Id. Table IV illustrates processor cache states for a directory based mechanism. See paragraph [0063].

As understood, the processor cache states of Hughes et al. do not manage load and performance information relating to a main storage and a cache storage. Thus, Hughes et al. does not determine which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

7. U.S. Patent No. 6,463,509 B1

U.S. Patent No. 6,463,509 B1 to Teoman et al. discloses an apparatus and a method for caching data I a storage device of a computer system. A relatively high-speed, intermediate-volume storage device is operated as a user-configurable cache. Requests to access a mass storage device such as a disk or tape are intercepted by a device driver that compares the access request against a directory of the contents of the user-configurable cache. If the user-configurable cache contains the data sought to be accessed, the access request is carried out in the user-configurable cache instead of being forwarded to the device driver for the target mass storage device. Because the user-cache is implemented using memory having a dramatically shorter access time than most mechanical mass storage devices, the access request is fulfilled much more quickly than if the originally intended mass storage device was accessed. Data is preloaded and responsively cached in the user-configurable cache memory based on user preferences.

As discussed in the present application at paragraphs [0009] and [0013]-[0014], the cache device is directly connected to an expansion bus (I/O bus) of a host computer. Data retrieved from various storage devices through different types of interfaces can be cached in a caching device of a host operating system by changing device drives and data transfer paths in the host operating system. In this method, to cache a large volume of data, extremely flexible caching is realized by using a disk of an extremely large capacity as a cache storage medium. However, two serious problems arise. The first is that support from the operating system becomes necessary to process data caching and retrieval from the caching device. The use target operating system must support the hierarchy of the device driver. The second problem is that the consumption of the bandwidth of the caching process-use expansion bus becomes excessive. There is the potential for the data that is to be cached

to be transferred two times by the expansion bus in accordance with the type of bus used. The first time is when the data is transferred to the host memory from the device accessing the storage data, and the second time is when the data is transferred to the cache device from the host memory. There is the potential for performance in a region where input/output is concentrated, such as multimedia streaming, to drop due to overloading of the host I/O bus.

The reference does not teach managing load and performance information relating to a main storage and a cache storage, or determining which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

8. International Patent Publication No. WO 03/017598

This reference discloses a SCSI-to-IP cache storage system that interconnects a host computing device or a storage unit to a switched packet network. The cache storage system includes a SCSI interface 40 that facilitates system communications with the host computing device or the storage unit, and an Ethernet interface 42 that allows the system to receive data from and send data to the Internet. The cache storage system further comprises a processing unit 44 that includes a processor 46, a memory 48, and a log disk 52 configured as a sequential access device. The log disk 52 caches data along with the memory resident in the processing unit 44, wherein the log disk and the memory 48 are configured as a two-level hierarchical cache.

As discussed in the present application at paragraphs [0011] and [0015], the device adapter can cache, in a disk connected to an access device, data retrieved from a storage system connected to an IP network. Thus, caching operations can be implemented in the device adapter without overloading the host expansion bus. The device is not given the function of conducting bandwidth control to precisely measure the load of the cache storage being used and the load of the interface used to retrieve data that is not cached, and the data retrieval destination is determined depending on whether or not the data is in the cache. The function of dynamically determining the data retrieval destination using the load status of the interface is necessary in the case of the retrieval of data characterized by real time processing, such as multimedia data.

The reference does not teach managing load and performance information relating to a main storage and a cache storage, or determining which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

9. Japanese Patent Publication No. JP 2001-051901

This reference discloses a cache storage device to improve a cache hit rate. The cache storage device 10 is integrated into a disk device 8 of an animation stream server 1. The cache storage device 10 has a cache memory 11 and a cache controller 12 for controlling the operation of the cache memory 11 with the access of data from a CPU 3 to a data holding device 9. On the basis of an SCSI message 20 sent from an SCSI adapter 6 through an SCSI bus 7, the cache controller 12 can switch the control mode of discharging operation of data held in the cache memory 11 for each of the data. As a control mode, a mode for inhibiting the discharge of specified data held in the cache memory 11 and a mode for discharging the data held in the cache memory 11 with an ordinary method are provided.

As discussed in the present application at paragraph [0012], the use of the network bandwidth in the SAN cannot be reduced because caching is conducted in the access-destination storage device. Thus, this method cannot solve the drop in data transfer performance in the SAN. Moreover, ordinarily the size of the cache is not a size sufficient to retain a large amount of data, and with multimedia data files accessed by a streaming server, the efficiency of the overall cache drops in a case where multiple host servers continuously and simultaneously access large-volume data files.

The reference does not teach managing load and performance information relating to a main storage and a cache storage, or determining which of the cache storage and the main storage to use as the storage from which the requested data will be retrieved, as recited in sole independent claim 1.

(f) In view of this petition, the Examiner is respectfully requested to issue a first Office Action at an early date.

Respectfully submitted,

Chun-Pok Leung
Reg. No. 41,405

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400
Fax: 415-576-0300
Attachments
RL:rl
60459849 v1

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 February 2003 (27.02.2003)

PCT

(10) International Publication Number
WO 03/017598 A1

(51) International Patent Classification[7]: H04L 12/56

(21) International Application Number: PCT/US02/26292

(22) International Filing Date: 15 August 2002 (15.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/312,471 15 August 2001 (15.08.2001) US

(71) Applicant: THE BOARD OF GOVERNORS FOR HIGHER EDUCATION, STATE OF RHODE ISLAND AND PROVIDENCE PLANTATIONS [US/US]; 301 Promenade Street, Providence, RI 02908 (US).

(72) Inventor: YANG, Qing; 81 West Wind Road, Wakefield, RI 02879 (US).

(74) Agents: O'SHEA, Patrick, J. et al.; Samuels, Gauthier & Stevens, LLP, Suite 3300, 225 Franklin Street, Boston, MA 02110 (US).

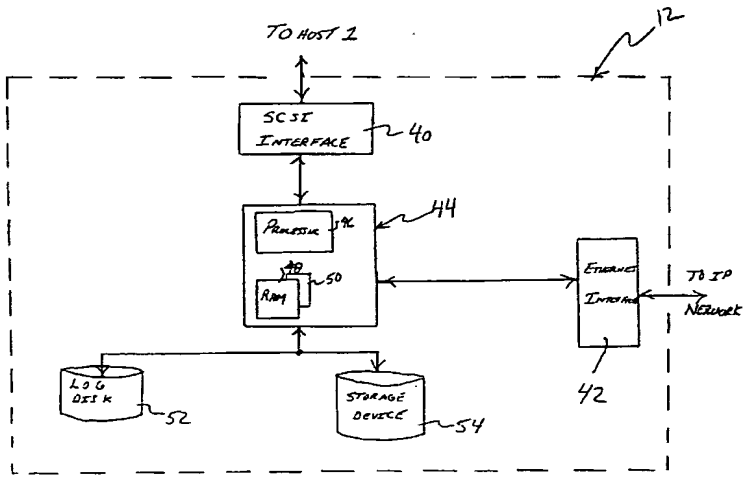
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SCSI-TO-IP CACHE STORAGE DEVICE AND METHOD



(57) Abstract: A SCSI-to-IP cache storage system interconnects a host computing device or a storage unit to a switched packet network. The cache storage system includes a SCSI interface (40) that facilitates system communications with the host computing device or the storage unit, and an Ethernet interface (42) that allows the system to receive data from and send data to the Internet. The cache storage system further comprises a processing unit (44) that includes a processor (46), a memory (48) and a log disk (52) configured as a sequential access device. The log disk (52) caches data along with the memory (48) resident in the processing unit (44), wherein the log disk (52) and the memory (48) are configured as a two-level hierarchical cache.

WO 03/017598 A1

SCSI-to-IP CACHE STORAGE DEVICE AND METHOD

PRIORITY INFORMATION

This application claims priority from provisional application Serial No. 60/312,471 filed August 15, 2002. This provisional application is hereby incorporated by reference.

BACKGROUND OF THE INVENTION

The invention relates to the field of data back-up systems, and in particular to a SCSI-to-IP cache storage device and method for implementing SAN over the Internet.

As we enter a new era of computing, data storage has changed its role from a secondary with respect to CPU and RAM, to a primary role in today's information world. Online data storage doubles approximately every nine months due to the increasing demand for network information services. While the importance of data storage is well-known, published literature is limited in the computer architecture research community reporting networked storage architecture. This situation will change quickly as information has surpassed raw computational power as the important commodity. This is especially true for Internet dependent businesses.

In general networked storage architectures have evolved from network-attached storage (NAS), storage area network (SAN), to more recent storage of IP (iSCSI). NAS architecture allows a storage system/device to be directly connected to a standard network, typically via the Ethernet. Clients on the network can access the NAS directly. A NAS based storage subsystem has a built-in file system to provide clients with file system functionality. SAN technology, on the other hand, provides a simple block level interface for manipulating nonvolatile magnetic media. Typically, a SAN includes networked storage devices interconnected through a dedicated Fibre Channel network. The basic premise of a SAN is to replace the current "point-to-point"

infrastructure with one that allows "any-to-any" communications. A SAN provides high connectivity, scalability, and availability using a specialized network interface - the Fibre Channel network. Deploying such a specialized network usually introduces cost for implementation, maintenance, and management. iCSI is the most recent emerging technology with the goal of implementing the SAN technology over the better-understood and mature network infrastructure, the Internet (TCP/IP).

Implementing SAN over IP brings economy and convenience whereas it also raises issues such as performance and reliability. Currently, there are basically two approaches: one encapsulates SCSI protocol in TCP/IP at host bus adapter (HBA) level, and the other carries out SCSI and IP protocol conversion at a specialized switch. However, both approaches have severe performance limitations. To encapsulate SCSI protocol over IP requires a significant amount of overhead traffic for SCSI command transfers and handshaking over the Internet. Converting protocols at a switch places a special burden on an already overloaded switch and creates another specialized piece of network equipment in the SAN. Furthermore, the Internet was not designed for transferring data storage blocks. Many features such as Maximum Transfer Unit (MTU), data gram fragmentation, routing, and congestion control may become obstacles to providing enough instant bandwidth for large block transfers of storage data.

Therefore, there is a need for a system that can implements SAN over switched packet network, such as for example the Internet (TCP/IP).

SUMMARY OF THE INVENTION

Briefly, according to an aspect of the invention, a SCSI-to-IP cache storage system includes a SCSI interface that facilitates system communication with host computers and extended storage devices. The system also includes an Ethernet interface that allows the system to receive data from and send data to the Internet, and a processing unit that includes a processor and memory. The system also includes a log disk that is a sequential access device. The log disk is used to cache data along with the memory resident in the processing unit. The log disk and the memory are configured as a two-level hierarchical cache for a disk storage device within the SCSI-to-IP cache storage system.

The system of the present invention facilitates implementing SAN over the Internet. The disk storage device within the SCSI-to-IP cache storage system is preferably configured as RAID.

Besides the regular data storage in the SCSI-to-IP cache storage system, one storage device within the system is used as a non-volatile cache that caches data coming from possibly two directions. That is, block data may come from the SCSI interface, and network data may come from the Ethernet interface. In addition, to standard SCSI and IP protocols running on the intelligent processing unit, a local file system may also reside in the processing unit. The file system is preferably a simplified Log-structured file system that writes data quickly and provides advantages to cache data both ways. Besides caching storage data in both directions, the SCSI-to-IP cache storage systems may also localize SCSI commands and handshaking operations to reduce unnecessary traffic over the Internet. In this way, the SCSI-to-IP cache storage system acts as a storage filter to discard a fraction of the data that would otherwise move across the

Internet, reducing the bottleneck imposed on limited Internet bandwidth and increasing storage data rate.

The system of the present invention provides an iSCSI network cache to smooth out the traffic and improve system performance. Such a cache or bridge is not only helpful but also necessary to a certain degree because of the different nature of SCSI and IP such as speed, data unit size, protocols and requirements. Wherever there is a speed disparity, cache of course helps. Analogous to cache memory used to cache memory data for a CPU, the SCSI-to-IP cache storage system is a cache storage used to cache networked storage data for a server host.

The system of the present invention may utilize the Log-structured file system to write data to magnetic media for caching data coming from both directions (e.g., from the Internet and from the host). In addition, since the SCSI-to-IP cache storage system preferably uses log disk to cache data, it is a nonvolatile cache, which is of course desirable for caching storage data reliably since once data is written to storage, it is considered safe.

The SCSI-to-IP cache storage system allows direct connection to a SCSI interface of a computer that in turn can access a SAN implemented over the Internet. In addition, by localizing part of the SCSI protocol and filtering of some unnecessary traffic, the SCSI-to-IP cache storage system can reduce the bandwidth requirement of the Internet to implement the SAN.

These and other objects, features and advantages of the present invention will become apparent in light of the following detailed description of preferred embodiments thereof, as illustrated in the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram illustration of a distributed computing and information backup

system that includes a SCSI-to-IP cache storage system;

FIG. 2 is a block diagram illustration of one the SCSI-to-IP cache storage systems illustrated in FIG. 1;

FIG. 3 is a block diagram illustration of a RAM buffer layout; and

FIG. 4 is a block diagram illustration of a SCSI initiator and target sub-systems.

DETAILED DESCRIPTION OF THE INVENTION

FIG. 1 is a block diagram illustration of a distributed computing and information backup system 10 that includes a plurality of SCSI-to-IP cache storage systems 12-15. Each of the plurality of SCSI-to-IP cache storage systems 12-15 interfaces one of an associated networked device such as a host 18, 20 or a storage device 22, 24 to the Internet. The distributed computing and information backup system 10 also includes network attached storage (NAS) 26 that communicates via the Internet. The system 10 provides a SAN implementation over IP using the SCSI-to-IP cache storage systems 12-15. Although the system 10 illustrated in FIG. 1 includes four SCSI-to-IP cache storage systems 12-15, one of ordinary skill will recognize that virtually any number of computing devices or storage devices can be connected through an associated SCSI-to-IP cache storage system to form the SAN. Significantly, rather than using a specialized network or storage switch, the SCSI-to-IP cache storage systems 12-15 connect a host computer or a storage device to the IP network. The SCSI-to-IP cache storage systems 12-15 each provide SCSI protocol service, caching service, naming service and protocol service.

FIG. 2 is a block diagram illustration of one of the SCSI-to-IP cache storage systems 12. This cache storage system 12 includes a SCSI interface 40 that supports SCSI communication with the host 18 (FIG. 1) and runs in a target mode receiving requests from the host 18 (FIG. 1),

carrying out the I/O processing through the network, and sending back results to the host 18 (FIG. 1). When the SCSI-to-IP cache storage system is used to connect a storage device such as a disk or RAID 22 (FIG. 1) to extended storage, the SCSI-to-IP cache storage system operates in an initiator mode, wherein it sends/forwards SCSI requests to the extended storage mode. Referring to FIG. 1, the SCSI-to-IP cache storage system 12 operates in target mode, while the SCSI-to-IP cache storage system 15 operates in initiator mode. The SCSI-to-IP cache storage system 12 acts as a directly attached local storage device to the host 18 (FIG. 1).

Referring still to FIG. 2, the SCSI-to-IP cache storage system 12 also includes an Ethernet interface 42, which connects the SCSI-to-IP cache storage system to the Internet. A processing unit 44 that includes a processor 46 and RAM 48 is also included in the cache storage system 12. The processing unit 44 executes a log-structured file system, SCSI protocol and IP protocols. The RAM 48 is primarily used as a buffer cache. NVRAM 50 (e.g., 1- 4 MB) is also included to maintain meta data such as a hash table, LRU list, and mapping information (STICS_MAP). The meta data is stored in the NVRAM 50 before being written to disk. This of course reduces the frequency of having to write or read meta data to/from disk. Alternatively, it is contemplated that Soft Updates as disclosed in the printed publication by G. Granger, M. McKusick, C. Soules and Y. Pratt, *"Soft Updates: A Solution to the Metadata Update Problem in File Systems"*, ACM Transactions on Computer Systems, Vol. 18, No. 2, 2000, pp. 127-135 may be used to keep meta data consistency without using NVRAM. This paper is hereby incorporated by reference.

The SCSI-to-IP cache storage system 12 further comprises a log disk 52, which is a sequential accessed device. The log disk is used to cache data along with the RAM within the processing unit 44. The log disk 52 and the RAM form a two-level hierarchical cache.

The system 12 also includes a storage device 54. The storage device 54 may be configured as a disk, a RAID, or just-bunch-of-disks (JBOD). The storage device 54 can be considered a local disk from the point of view of the host 18 (FIG. 1). From the point of view of the IP network via the network interface 42, the storage device 54 is considered as a component of a networked storage system such as a SAN with an IP address as its ID.

To allow a true "any-to-any" communication between servers and storage devices, a global naming is required. In one embodiment, each of the SCSI-to-IP cache storage systems 12-15 (FIG. 1) is named by a global location number (GLN) which is unique for each of the SCSI-to-IP cache storage systems. An IP address is assigned to each SCSI-to-IP cache storage system and use this IP as the GLN.

The cache organization in the SCSI-to-IP cache storage system includes a two level hierarchy: a RAM cache and a log disk. Frequently accessed data reside in the RAM, which is organized as a LRU cache 58 as shown in FIG. 3. Whenever the newly written data in the RAM are sufficiently large or whenever the log disk 52 (FIG. 2) is free, data are written into the log disk. There are also less frequently accessed data kept in the log disk. Data in the log disk are organized in the format of *segments* similar to that in a Log-structured File System. A segment contains a plurality of *slots* each of which can hold one data block. Data blocks in segments are addressed by their *Segment* IDs and *Slot IDs.*

One of the challenging tasks in this research is to design an efficient data structure and a search algorithm for RAM cache. As shown in FIG. 3, the RAM cache includes a hash table that is used to locate data in the cache, a data buffer which contains several data slots, and a few In-memory headers. Data blocks stored in the RAM cache are addressed by their *Logical Block Addresses (LBAs).* The hash table contains location information for each of the valid data blocks

in the cache and uses LBAs of incoming requests as search keys. The slot size is set to be the size of a block. A slot entry includes the following fields:

- An LBA entry that is the LBA of the cache line and serves as the search key of hash table;
- Global Location Number (GLN) if the slot contains data from or to other STICS.
- A log disk LBA is divided into at least two parts:

1. A state tag (2 bits), used to specify where the slot data is: IN_RAM_BUFFER, IN_LOG_DISK, IN_DATA_DISK or IN_OTHER_STICS;

2. A log disk block index (e.g., 30 bits), used to specify the log disk block number if the state tag indicates IN_LOG_DISK. The size of each log disk can be up to for example 2^{30} blocks.

- Two pointers (hash_prev and hash_next) are used to link the hash table;
- Two pointers (prev and next) are used to link the LRU list and FREE list;
- A Slot-No is used to describe the in-memory location of the cached data.

As set forth above, the SCSI-to-IP cache storage system may run under two modes: (i) initiator mode or (ii) target mode. FIG. 4 is a block diagram illustration of the SCSI-to-IP cache storage system initiator 100 and target modes 110. When running in target mode, the SCSI-to-IP cache storage system is connected to a host and the host is running in initiator mode. Otherwise the SCSI-to-IP cache storage system runs in initiator mode. Initiator mode is the default mode of the SCSI-to-IP cache storage system. All server host platforms including Linux support SCSI initiator mode. The standard SCSI initiator mode operates in the SCSI-to-IP cache storage system. The SCSI target runs in parallel to the initiator and is concerned only with the processing of SCSI commands. A set of target APIs is defined for the SCSI-to-IP cache storage

system. These APIs include SCSI functions such as SCSI_DETECT, SCSI_RELEASE, SCSI_READ, SCSI_WRITE and etc. When running under target mode, a SCSI-to-IP cache storage system looks like a standard SCSI device to a connected host.

For each the SCSI-to-IP cache storage system, a variable STICS_LOAD is defined to represent its current load. The higher the STICS_LOAD, the busier the SCSI-to-IP cache storage system is. When a SCSI-to-IP cache storage system starts, its STICS_LOAD is set to zero. When the SCSI-to-IP cache storage system accepts a request, STICS_LOAD is decremented. Besides STICS_LOAD, STICS_MAP is defined to map all the SCSI-to-IP cache storage system loads within the network. STICS_MAP is a set of <GLN, STICS_LOAD> pairs. The STICS_MAP is also updated dynamically.

Write requests may come from one of two sources: the host via the SCSI interface and from another SCSI-to-IP cache storage system via the Ethernet interface. The operations of these two types of writes are as follows.

After receiving a write request from the host via the SCSI interface, the SCSI-to-IP cache storage system searches the hash table by the LBA address. If an entry is found, the entry is overwritten by the incoming write. Otherwise, a free slot entry is allocated from the Free List, the data are copied into the corresponding slot, and its address is recorded in the hash table. The LRU list and Free List are then updated. When enough data slots (e.g., sixteen) are accumulated or when the log disk is idle, the data slots are written into log disk sequentially in one large write. After the log write completes successfully, the SCSI-to-IP cache storage system signals the host that the request is complete.

A packet coming from another the SCSI-to-IP cache storage system via the Ethernet interface may turn out to be a write operation from a remote SCSI-to-IP cache storage system on

the network. After receiving such a write request and unpacking the network packet, SCSI-to-IP cache storage systems gets a data block with GLN and LBA. It then searches the Hash Table by the LBA and GLN. If an entry is found, the entry is overwritten by the incoming write. Otherwise, a free slot entry is allocated from the Free List, and the data are then copied into the corresponding slot. Its address is recorded in the Hash Table. The LRU list and Free List are updated accordingly.

Similar to write operations, read operations may also come either from the host via the SCSI interface or from another SCSI-to-IP cache storage system via the Ethernet interface.

After receiving a read request from the host via the SCSI interface, the SCSI-to-IP cache storage system searches the Hash Table by the LBA to determine the location of the data. Data requested may be in one of four different places: (i) the RAM buffer, (ii) the log disk(s), (iii) the storage device in the local SCSI-to-IP cache storage system, or (iv) a storage device in another SCSI-to-IP cache storage system on the network. If the data is found in the RAM buffer, the data are copied from the RAM buffer to the requesting buffer. The SCSI-to-IP cache storage system then signals the host that the request is complete. If the data is found in the log disk or the local storage device, the data are read from the log disk or storage device into the requesting buffer. Otherwise, the SCSI-to-IP cache storage system encapsulates the request including LBA, current GLN, and destination GLN into an IP packet and forwards it to the corresponding SCSI-to-IP cache storage system.

When a read request from another SCSI-to-IP cache storage system via the Ethernet interface is found after unpacking an incoming IP packet, the SCSI-to-IP cache storage system obtains the GLN and LBA from the packet. It then searches the Hash Table by the LBA and the source GLN to determine the location of the data. It locates and reads the data from that

location. It sends the data back to the source SCSI-to-IP cache storage system through the network.

The operation of moving data from a higher-level storage device to a lower level storage device is defined as *destage* operation. There are two levels of destage operations in the SCSI-to-IP cache storage systems: (i) destaging data from the RAM buffer to the log disk *(level 1 destage)* and (ii) destaging data from log disk to a storage device (*level 2 destage)*. A separate kernel thread, *LogDestage*, is implemented to perform the destaging tasks. The *LogDestage* thread is registered during system initialization and monitors the SCSI-to-IP cache storage system states. The thread remains asleep most of the time, and is activated when one of the following events occurs: (i) the number of slots in the RAM buffer exceeds a threshold value, (ii) the log disk is idle, (iii) the SCSI-to-IP cache storage system detects an idle period, or (iv) the SCSI-to-IP cache storage system RAM buffer and/or the log disk becomes full. *Level 1 Destage* has higher priority than *Level 2 Destage*. Once the *Level 1 destage* starts, it continues until a log of data in the RAM buffer is written to the log disk. *Level 2 destage* may be interrupted if a new request comes in or until the log disk becomes empty. If the destage process is interrupted, the destage thread is suspended until the SCSI-to-IP cache storage system STICS detects another idle period.

For *Level 1 Destage*, the data in the RAM buffer are written to the log disk sequentially in large size (e.g., 63KB). The log disk header and the corresponding in-memory slot entries are updated. All data are written to the log disk in "append" mode, which insures that every time the data are written to consecutive log disk blocks.

For *Level 2 destage*, a "last-write-first-destage" algorithm is employed according to the LRU List. At this point, a SCSI-to-IP cache storage system with the lowest STICS_LOAD is

selected to accept data. Each time 64KB data are read from the consecutive blocks of the log disk and written to the chosen SCSI-to-IP cache storage system storage disks. The LRU list and free list are updated subsequently.

Advantageously, the SCSI-to-IP cache storage system facilitates implementation of SAN over the Internet. The SCSI-to-IP cache storage system allows any server host to access a SAN on Internet through a standard SCSI interface. Using a non-volatile "cache storage", the SCSI-to-IP cache storage system smoothes out the storage data traffic between SCSI and IP, analogous to the way the cache memory smoothes out CPU-memory traffic.

Although the present invention has been shown and described with respect to several preferred embodiments thereof, various changes, omissions and additions to the form and detail thereof, may be made therein, without departing from the spirit and scope of the invention.

What is claimed is:

CLAIMS

1. A SCSI-to-IP cache storage system that interconnects a host computing device or a storage unit to a switched packet network, said cache storage system comprising:

a SCSI interface that facilitates system communication with the host computing device or the storage unit;

an Ethernet interface that allows the system to receive data from and send data to the Internet;

a processing unit that includes a processor and memory;

a log disk configured as a sequential access device that caches data along with said memory resident in said processing unit, wherein said log disk and said memory are configured as a two-level hierarchical cache; and

a disk storage device that receives data from and provides data to the two level hierarchical cache.

2. The SCSI-to-IP cache storage system of claim 1, wherein said memory comprises RAM.

3. The SCSI-to-IP cache storage system of claim 2, wherein said RAM comprises NVRAM.

4. The SCSI-to-IP cache storage system of claim 3, wherein said disk storage device comprises a redundant array of inexpensive disks.

5. A SCSI-to-IP cache storage system that interconnects a host computing device or a

storage unit to a switched packet network, said cache storage system comprising:

a SCSI interface that facilitates system communication with the host computing device or the storage unit;

an Ethernet interface that allows the system to receive data from and send data to the Internet;

a processing unit that includes a processor and RAM;

a log disk that caches data along with said RAM, wherein said log disk and said RAM are configured as a two-level hierarchical cache; and

a disk storage device that receives data from and provides data to the two level hierarchical cache.

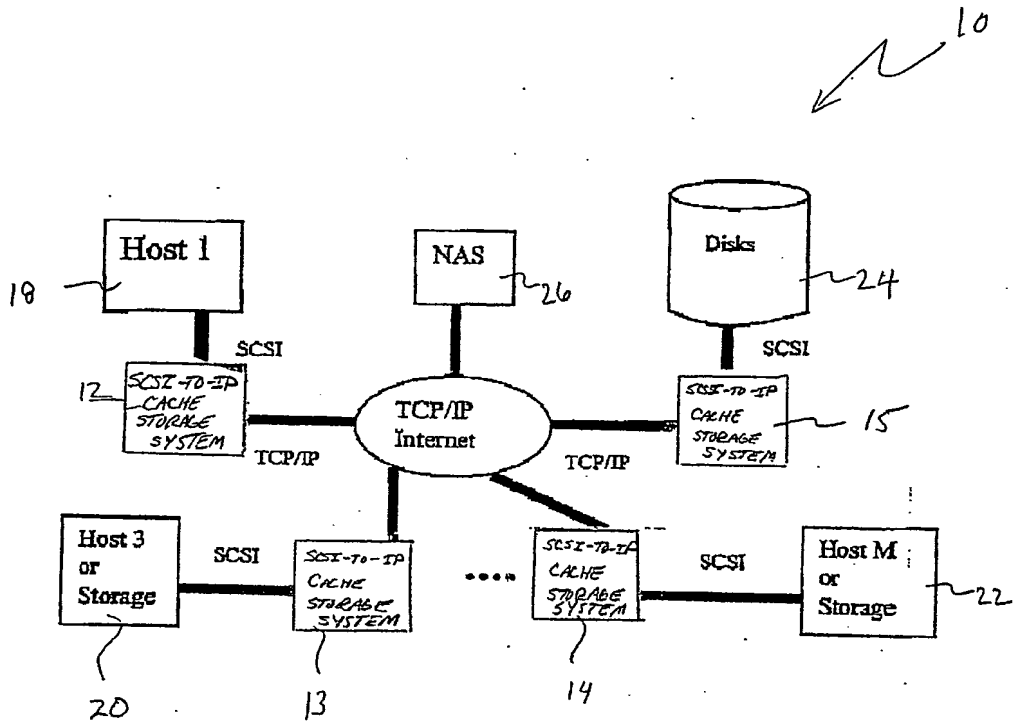

10
Host 1
18
NAS
26
Disks
24
SCSI
SCSI
12
SCSI-TO-IP CACHE STORAGE SYSTEM
TCP/IP Internet
TCP/IP
TCP/IP
SCSI-TO-IP CACHE STORAGE SYSTEM
15
Host 3 or Storage
SCSI
SCSI-TO-IP CACHE STORAGE SYSTEM
SCSI-TO-IP CACHE STORAGE SYSTEM
SCSI
Host M or Storage
22
20
13
14

FIG. 1

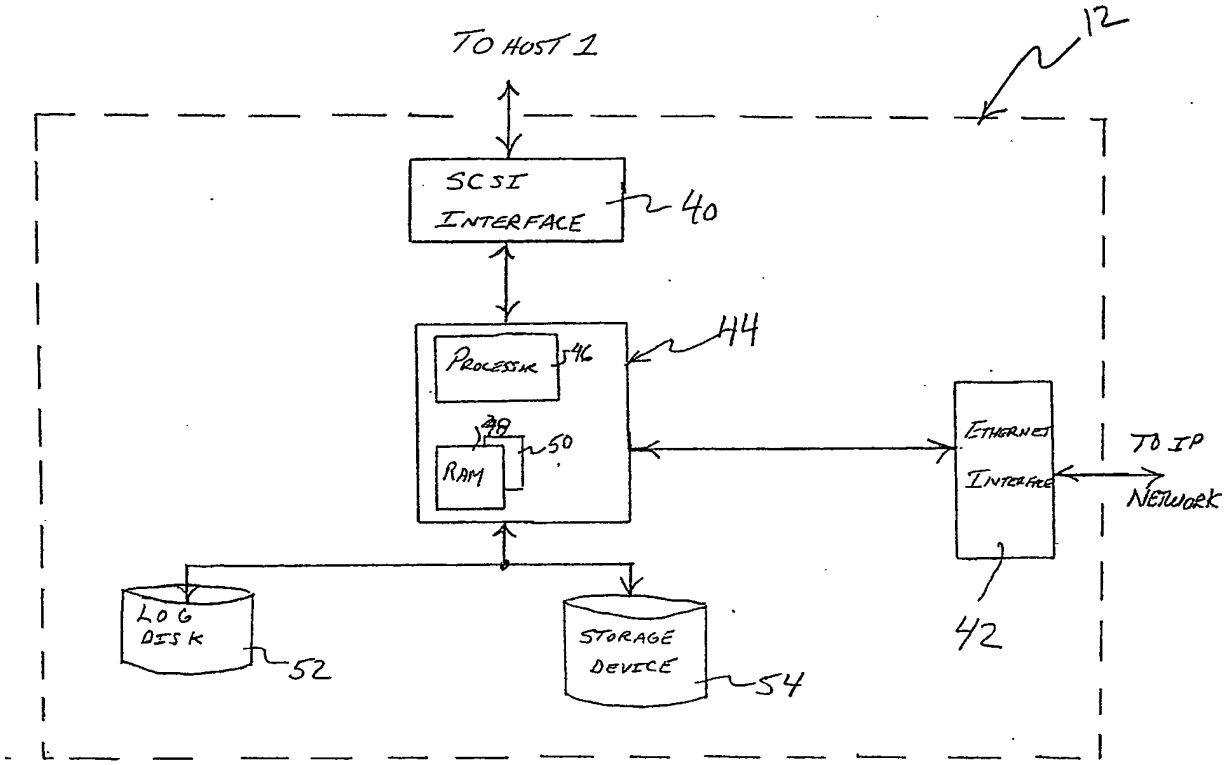

TO HOST 1
12
SCSI INTERFACE
40
PROCESSOR
46
44
48
RAM
50
ETHERNET INTERFACE
TO IP NETWORK
42
LOG DISK
52
STORAGE DEVICE
54

FIG. 2

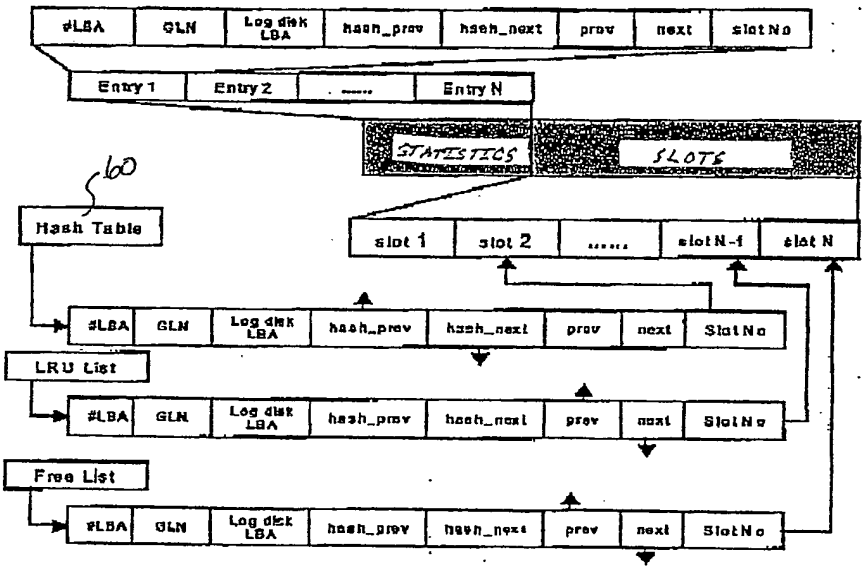

#LBA
GLN
Log disk LBA
hash_prev
hash_next
prev
next
slot No
Entry 1
Entry 2
Entry N
STATISTICS
SLOTS
60
Hash Table
slot 1
slot 2
slot N-1
slot N
LRU List
Free List

FIG. 3

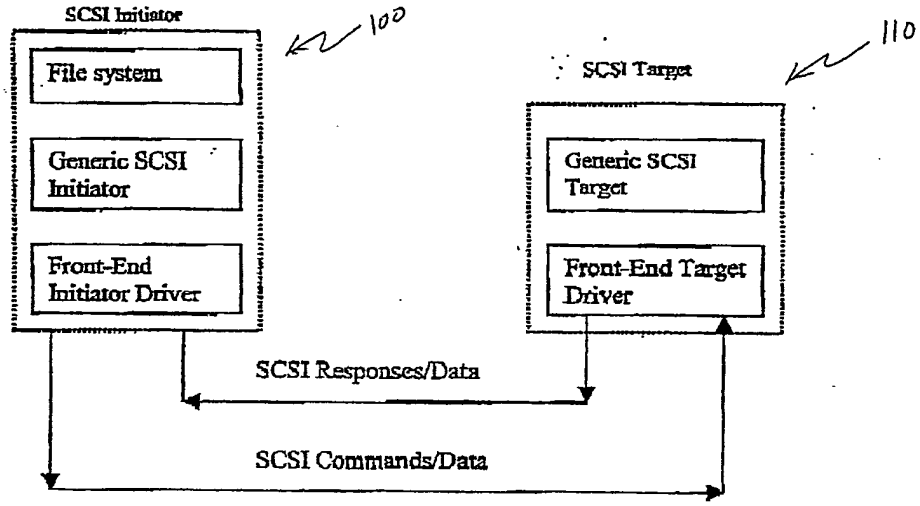

SCSI Initiator
100
File system
Generic SCSI Initiator
Front-End Initiator Driver
SCSI Target
110
Generic SCSI Target
Front-End Target Driver
SCSI Responses/Data
SCSI Commands/Data

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US02/26292

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : ,H04L 12/56
US CL : 370/395.72; 709/213
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
U.S. : 370/395.72, 363, 368, 371, 374, 378, 381, 383, 395.7, 395.71, 428; 709/213, 214, 215, 216, 217, 218, 219, 312, 321, 325, 326

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EAST, NPL

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PRESLAN, K. W. et al. A 64-Bit, Shared Disk File System for Linux. Mass Storage Systems, 15-18 March 1999, pages 22-41.	1-5
A	VORUGANTI, K. et al. An Analysis of Three Gigabit Networking Protocols for Storage Area Networks. Performance, Computing, and Communications, 4-6 April 2001, pages 259-265.	1-5
A	LEE, Y. K. et al. Metadata Management of the SANtopia File System. Parallel and Distributed Systems, 26-29 June 2001, pages 492-499.	1-5
A	KIM, C. S. et al. Volume Management in SAN Environment. Parallel and Distributed Systems, 26-29 June 2001, pages 500-505.	1-5
A	MOLERO, X. et al. On the Switch Architecture for Fibre Channel Storage Area Networks. Parallel and Distributed Systems, 26-29 June 2001, pages 484-491.	1-5
A	NAMGOONG, J. C. et al. Design and Implementation of a Fibre Channel Network Driver for SAN-attached RAID controllers. Parallel and Distributed Systems, 26-29 June 2001, pages 477-483.	1-5

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"B" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
25 October 2002 (25.10.2002)

Date of mailing of the international search report
06 DEC 2002

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703)305-3230

Authorized officer
Hassan Kizou
Telephone No. 703-305-4730

Form PCT/ISA/210 (second sheet) (July 1998)

PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2001-051901**
(43)Date of publication of application : **23.02.2001**

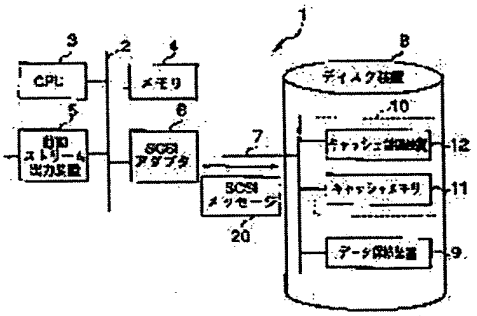
(51)Int.Cl. **G06F 12/12**
G06F 3/06
G06F 12/00
G06F 12/08

(21)Application number : **11-223031**
(22)Date of filing : **05.08.1999**

(71)Applicant : **TOSHIBA CORP**
(72)Inventor : **MIURA MASAHIRO**

(54) CACHE STORAGE DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To improve a cache hit rate by surely remaining data expected to be frequently accessed in a cache memory.
SOLUTION: This device 10 is integrated into a disk device 8 of an animation stream server 1. The cache storage device 10 has a cache memory 11 and a cache controller 12 for controlling the operation of the cache memory 11 with the access of data from a CPU 3 to a data holding device 9. On the basis of an SCSI message 20 sent from an SCSI adapter 6 through an SCSI bus 7, the cache controller 12 can switch the control mode of discharging operation of data held in the cache memory 11 for each of data. As a control mode, a mode for inhibiting the discharge of specified data held in the cache memory 11 and a mode for discharging the data held in the cache memory 11 with an ordinary method are provided.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－51901
(P2001－51901A)
(43)公開日　平成13年２月23日(2001.2.23)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'(参考)
Ｇ０６Ｆ　12/12		Ｇ０６Ｆ　12/12		Ｆ　５Ｂ００５
				Ａ　５Ｂ０６５
				Ｄ　５Ｂ０８２
3/06	３０２		3/06	３０２Ａ
12/00	５１４		12/00	５１４Ｍ

審査請求　未請求　請求項の数４　ＯＬ　（全６頁）　最終頁に続く

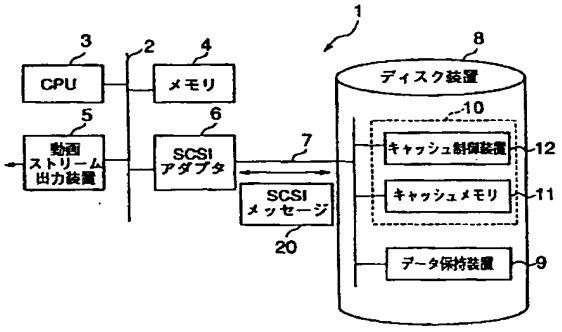
(21)出願番号　特願平11－223031
(22)出願日　平成11年８月５日(1999.8.5)

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　三　浦　雅　弘
東京都府中市東芝町１番地　株式会社東芝
府中工場内
(74)代理人　100064285
弁理士　佐藤　一雄　（外３名）

Ｆターム(参考)　5B005 JJ13 MM03 MM04 NN43 QQ05
5B065 BA01 CA16 CE12 CH03
5B082 EA07 FA12

(54)【発明の名称】　キャッシュ記憶装置

(57)【要約】
【課題】　頻繁にアクセスされると予想されるデータが確実にキャッシュメモリに残るようにしてキャッシュヒット率の向上を図ることができるキャッシュ記憶装置を提供する。
【解決手段】　キャッシュ記憶装置１０は、動画ストリームサーバ１のディスク装置８内に組み込まれている。キャッシュ記憶装置１０は、キャッシュメモリ１１と、ＣＰＵ３からデータ保持装置９へのデータのアクセスに伴うキャッシュメモリ１１の動作を制御するキャッシュ制御装置１２とを有している。キャッシュ制御装置１２は、ＳＣＳＩアダプタ６からＳＣＳＩバス７を介して送られてきたＳＣＳＩメッセージ２０に基づいて、キャッシュメモリ１１で保持されるデータの追い出し動作の制御モードをデータごとに切り替えることができる。制御モードとしては、キャッシュメモリ１１で保持される特定のデータの追い出しを禁止するモードと、キャッシュメモリ１１で保持されるデータを通常の方法で追い出すモードとが設けられている。



2003 11 19 09:45

【特許請求の範囲】
【請求項１】中央処理装置と主記憶装置との間に設けられ前記主記憶装置に格納されたデータの一部を保持するキャッシュメモリと、
前記中央処理装置から前記主記憶装置へのデータのアクセスに伴う前記キャッシュメモリの動作を制御するキャッシュ制御装置とを備え、
前記キャッシュ制御装置は、外部から与えられた指示に基づいて、前記キャッシュメモリで保持されるデータの追い出し動作の制御モードをデータごとに切り替えることを特徴とするキャッシュ記憶装置。
【請求項２】前記制御モードの一つとして、前記キャッシュメモリで保持される特定のデータの追い出しを禁止するモードを含むことを特徴とする請求項１記載のキャッシュ記憶装置。
【請求項３】前記制御モードの一つとして、前記キャッシュメモリで保持される特定のデータを強制的に追い出すモードを含むことを特徴とする請求項１記載のキャッシュ記憶装置。
【請求項４】前記キャッシュ制御装置は、前記キャッシュメモリで保持される複数のデータブロックを管理する管理テーブルであって前記各データブロックの追い出し動作の制御モードを表す管理フラグが設けられた管理テーブルを有し、この管理テーブルに設けられた管理フラグを外部から与えられた指示に基づいて設定することにより、キャッシュメモリで保持される各データブロックの追い出し動作の制御モードをデータブロックごとに切り替えることを特徴とする請求項１記載のキャッシュ記憶装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は主記憶装置に格納されたデータへの高速なアクセスを実現するキャッシュ記憶装置に係り、とりわけキャッシュヒット率の向上を図ることができるキャッシュ記憶装置に関する。
【０００２】
【従来の技術】従来から、大容量の主記憶装置（メモリやディスク装置等）へのアクセス回数を減少させることによってデータ処理速度を向上させる手法として、中央処理装置と主記憶装置との間にキャッシュメモリを設ける手法が知られている。ここで「キャッシュメモリ」とは、主記憶装置へのアクセスで得られたデータを一時的に保持しておく高速なメモリであり、以降の同一のデータへのアクセスをこの高速なメモリに対して行うことによりデータ処理速度を向上させることができる。なお、キャッシュメモリでデータを保持することを通常「キャッシュする」という。
【０００３】ところで、このようなキャッシュメモリは比較的小容量のメモリであるので、主記憶装置へのアクセスで得られたデータの全てをキャッシュしておくことはできない。また、当然ながら、初めてアクセスするデータはキャッシュされていない。このため、アクセスしようとするデータが常にキャッシュされているとは限らないことになる。ここで、アクセスしようとするデータがキャッシュされていることを「キャッシュヒット」といい、キャッシュヒットの起こる割合のことを「キャッシュヒット率」という。
【０００４】キャッシュメモリを用いたデータ処理システムにおいて、データ処理速度をより向上させるためには、当然ながら、キャッシュヒット率を向上させる必要がある。上述したように、キャッシュメモリは比較的小容量のメモリであり、主記憶装置へのアクセスで得られたデータを全てキャッシュしておくことはできないので、キャッシュメモリが一杯になった場合にはある適当な方法でデータを選んでキャッシュから追い出す、という処理（以下「追い出し動作」という）を行う必要がある。ここで、キャッシュヒット率とは、アクセスしようとするデータがどの程度キャッシュメモリに残っているかということに依存する指標であるので、キャッシュヒット率を向上させるためには、上述したような追い出し動作において、近い将来再びアクセスされるデータをキャッシュメモリに残し、もうアクセスされなくなったデータをキャッシュメモリから追い出す必要がある。
【０００５】従来においては、ディスク装置内や中央処理装置内等に組み込まれたキャッシュ制御装置により、「アクセスされたのが古いデータほど、アクセスされない可能性が高い」といった経験的なアルゴリズムに基づいて、ＦＩＦＯ（First-In First-Out）法や、ＬＲＵ（Least Recently Used）法といった決定法により、追い出すべきデータを選ぶというのが一般的である。
【０００６】
【発明が解決しようとする課題】しかしながら、あるデータへのアクセスが将来どのようになるのかということを予測することは非常に困難である。具体的には例えば、上述した従来の方法では、ある程度大きいファイルを繰り返してアクセスするような場合、そのファイル自体の再利用特性等を考慮することなく単純に個々のデータへのアクセス状況等に基づいて追い出し動作を行うので、キャッシュメモリから追い出されたデータが直後にまたアクセスされる、というような事態が生じやすい。
【０００７】本発明はこのような点を考慮してなされたものであり、頻繁にアクセスされると予想されるデータが確実にキャッシュメモリに残るようにしてキャッシュヒット率の向上を図ることができるキャッシュ記憶装置を提供することを目的とする。
【０００８】
【課題を解決するための手段】本発明は、中央処理装置と主記憶装置との間に設けられ前記主記憶装置に格納されたデータの一部を保持するキャッシュメモリと、前記中央処理装置から前記主記憶装置へのデータのアクセス

に伴う前記キャッシュメモリの動作を制御するキャッシュ制御装置とを備え、前記キャッシュ制御装置は、外部から与えられた指示に基づいて、前記キャッシュメモリで保持されるデータの追い出し動作の制御モードをデータごとに切り替えることを特徴とするキャッシュ記憶装置である。

【０００９】なお、本発明においては、前記制御モードの一つとして、前記キャッシュメモリで保持される特定のデータの追い出しを禁止するモードを含むことが好ましい。また、前記制御モードの一つとして、前記キャッシュメモリで保持される特定のデータを強制的に追い出すモードを含むことが好ましい。さらに、前記キャッシュ制御装置は、前記キャッシュメモリで保持される複数のデータブロックを管理する管理テーブルであって前記各データブロックの追い出し動作の制御モードを表す管理フラグが設けられた管理テーブルを有し、この管理テーブルに設けられた管理フラグを外部から与えられた指示に基づいて設定することにより、キャッシュメモリで保持される各データブロックの追い出し動作の制御モードをデータブロックごとに切り替えることが好ましい。

【００１０】本発明によれば、キャッシュメモリで保持されるデータに対して外部から適切な指示を与えることにより、頻繁にアクセスされると予想されるデータを確実にキャッシュメモリに残すことができるので、キャッシュヒット率を向上させることができ、全体のデータ処理性能を高めることができる。

【００１１】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態について説明する。図１乃至図３は本発明によるキャッシュ記憶装置の一実施の形態を説明するための図である。なお、本実施の形態では、キャッシュ記憶装置が動画ストリームサーバのディスク装置内に組み込まれている場合を例に挙げて説明する。

【００１２】まず、図１により、キャッシュ記憶装置が適用される動画ストリームサーバの全体構成について説明する。

【００１３】図１に示すように、動画ストリームサーバ１は、メインバス２を備え、このメインバス２にはＣＰＵ（中央処理装置）３、メモリ４、動画ストリーム出力装置５およびＳＣＳＩアダプタ６が接続されている。ＳＣＳＩアダプタ６にはＳＣＳＩバス７を介してディスク装置８が接続されており、ＳＣＳＩアダプタ６とディスク装置８との間でＳＣＳＩメッセージ２０がやりとりされるようになっている。また、ディスク装置８は、磁気ディスク装置等からなるデータ保持装置（主記憶装置）９と、データ保持装置９に格納されたデータへの高速なアクセスを実現するキャッシュ記憶装置１０とを備えている。なお、データ保持装置９には、外部からの要求に応じて動画ストリーム出力装置５を介して出力される動画ストリームのデータが格納されている。

【００１４】ここで、キャッシュ記憶装置１０は、キャッシュメモリ１１と、ＣＰＵ３からデータ保持装置９へのデータのアクセス（データの読み出しまたは書き込み）に伴うキャッシュメモリ１１の動作を制御するキャッシュ制御装置１２とを有している。キャッシュ制御装置１２は、ＳＣＳＩアダプタ６からＳＣＳＩバス７を介して送られてきたＳＣＳＩメッセージ２０に基づいて、キャッシュメモリ１１で保持されるデータの追い出し動作の制御モードをデータごとに切り替えることができるようになっている。なお、本実施の形態では、制御モードとして、キャッシュメモリ１１で保持される特定のデータの追い出しを禁止するモード（禁止モード）と、キャッシュメモリ１１で保持されるデータを通常の方法で追い出すモード（通常モード）とが設けられているものとする。なお、制御モードは、この２つに限られるものではなく、データの追い出しを極力禁止するといった禁止モードと通常モードとの中間的なモードを設けるようにしてもよい。

【００１５】図２はキャッシュ記憶装置１０の論理構成を示す図であり、符号１３はキャッシュ制御装置１２で保持される管理テーブルを示している。管理テーブル１３は、キャッシュメモリ１１で保持される複数のデータブロック１１ａを管理するためのものであり、管理フラグ１３ａ、最近アクセス時刻１３ｂおよびポインタ１３ｃという３つのエントリを有している。このうち、管理フラグ１３は、各データブロック１１ａの追い出し動作の制御モード（例えば値“１”は禁止モード、値“０”は通常モード）を表しており、この値に基づいて、キャッシュ制御装置１２により、キャッシュメモリ１１で保持される各データブロック１１ａの追い出し動作の制御モードをデータブロックごとに切り替えることができるようになっている。なお、管理フラグ１３ａは、ＳＣＳＩアダプタ６からＳＣＳＩバス７を介して送られてきたＳＣＳＩメッセージ２０に基づいて設定されるようになっている。

【００１６】次に、このような構成からなる本実施の形態の作用について説明する。

【００１７】まず、動画ストリームサーバ１において、外部からの要求に応じて、今後頻繁に出力されると予想される動画ストリーム（例えば、クライアントからの要求に応じて随時出力される人気番組等の動画ストリーム）を出力する場合を想定する。この場合には、ディスク装置８のデータ保持装置９から読み出される動画ストリームのデータは、そのサイズ、キャッシュメモリ１１のサイズ、および出力頻度等によっては、キャッシュメモリ１１で常時保持しておくとキャッシュヒット率が向上する。

【００１８】そこで、ＣＰＵ３は、このような動画ストリームのデータの読み出しが外部から要求された場合に、データの読み出しに先立ってＳＣＳＩアダプタ６お

よびＳＣＳＩバス７を介して図３（ａ）に示すようなＳＣＳＩメッセージ２０をディスク装置８内のキャッシュ制御装置１２に送り、このＳＣＳＩメッセージ２０に続いて読み出されるデータを禁止モードでキャッシュする。なお、図３（ａ）において、符号２１は拡張メッセージであることを表すメッセージコード（“０１ｈ”）を示し、符号２２は以降のメッセージバイト数（“０２ｈ”）を示し、符号２３はベンダ拡張コード（“？？ｈ”（８０ｈ～ｆｆｈ））を示し、符号２４はデータを禁止モードでキャッシュすることを表すフラグ（“０１ｈ”）を示している。

【００１９】このとき、キャッシュ制御装置１２は、図３（ａ）に示すＳＣＳＩメッセージ２０に続いて読み出されたデータをキャッシュメモリ１１で保持し、同時にそれらのデータに対応する管理フラグ１３ａを“１”に設定する。これにより、このようにしてキャッシュメモリ１１で保持されるデータは、キャッシュメモリ１１からの追い出しが禁止されることとなり、キャッシュメモリ１１で常時保持される。このため、今後頻繁に出力されると予想される動画ストリームのデータが確実にキャッシュメモリ１１で残るようになり、キャッシュヒット率を向上させることができる。

【００２０】なお、図３（ａ）に示すＳＣＳＩメッセージ２０が一度ディスク装置８内のキャッシュ制御装置１２に送られると、それ以後に読み出されるデータが常時禁止モードでキャッシュされることとなる。この状態を解消しようとする場合には、例えば、図３（ｃ）に示すようなＳＣＳＩメッセージ２０をディスク装置８内のキャッシュ制御装置１２に送るようにするとよい。なお、図３（ｃ）において、符号２１～２３は、図３（ａ）に示す内容と同一のものを表しており、符号２４は、データを通常モードでキャッシュすることを示すフラグ（“０２ｈ”）を示している。

【００２１】次に、動画ストリームサーバ１において、キャッシュメモリ１１で常時保持しておくよう指示された動画ストリームのデータが、番組の入れ替え等によりこれ以上出力されなくなった場合を想定する。この場合には、このような動画ストリームのデータをキャッシュメモリ１１で保持し続ける必要はない。

【００２２】そこで、ＣＰＵ３は、ＳＣＳＩアダプタ６およびＳＣＳＩバス７を介して図３（ｂ）に示すようなＳＣＳＩメッセージ２０をディスク装置８内のキャッシュ制御装置１２に送り、このＳＣＳＩメッセージ２０で指定されたデータを通常の方法で追い出すようにする。なお、図３（ｂ）において、符号２１～２３は、図３（ａ）に示す内容と同一のものを表しており、符号２４はデータの追い出しを許可することを示すフラグ（“０２ｈ”）を示し、符号２５は追い出しを許可するデータの場所（“Ａｈ”，“Ｂｈ”，…，“Ｃｈ”）を示している。

【００２３】このとき、キャッシュ制御装置１２は、図３（ｂ）に示すＳＣＳＩメッセージ２０で指定されたデータに対応する管理フラグ１３ａを“０”に設定する。これにより、このようにして指定されたデータは、キャッシュメモリ１１からの追い出しが許可されることとなり、キャッシュメモリ１１から通常の追い出し動作によって追い出しが行われる。

【００２４】最後に、動画ストリームサーバ１において、外部からの要求に応じて、一度出力すれば当分の間は出力する必要がないと予想される動画ストリーム（例えば、放送時刻の決まっている番組のオープニング等の動画ストリーム）を出力する場合を想定する。この場合には、このような動画ストリームのデータは、ディスク装置８のデータ保持装置９から当分読み出されないということであるので、キャッシュメモリ１１で保持しておく必要はない。

【００２５】そこで、ＣＰＵ３は、ＳＣＳＩアダプタ６およびＳＣＳＩバス７を介して図３（ｄ）に示すようなＳＣＳＩメッセージ２０をディスク装置８内のキャッシュ制御装置１２に送り、このＳＣＳＩメッセージ２０に続いて読み出されるデータをキャッシュしないようにする。これにより、キャッシュする必要のないデータがキャッシュメモリ１１で保持されることがなくなり、キャッシュヒット率を向上させることができる。なお、図３（ｄ）において、符号２１～２３は、図３（ａ）に示す内容と同一のものを表しており、符号２４は、データをキャッシュしないことを示すフラグ（“００ｈ”）を示している。

【００２６】なお、図３（ｄ）に示すＳＣＳＩメッセージ２０が一度ディスク装置８内のキャッシュ制御装置１２に送られると、それ以後に読み出されるデータが常時キャッシュされないようになる。この状態を解消しようとする場合には、例えば、上述した場合と同様に、図３（ｃ）に示すようなＳＣＳＩメッセージ２０をディスク装置８内のキャッシュ制御装置１２に送るようにするとよい。

【００２７】このように本実施の形態によれば、キャッシュメモリ１１で保持されるデータに対してＳＣＳＩメッセージ２０にて外部から適切な指示を与えることにより、頻繁にアクセスされると予想されるデータを確実にキャッシュメモリ１１に残すことができるので、キャッシュヒット率を向上させることができ、全体のデータ処理性能を高めることができる。

【００２８】なお、上述した実施の形態においては、ＳＣＳＩアダプタ６およびＳＣＳＩバス７を介して送られてきたＳＣＳＩメッセージ２０に基づいて、キャッシュメモリ１１で保持されるデータの追い出し動作の制御モードを切り替えるようにしているが、これに限らず、例えば個々のデータのアクセスの度にそのデータの追い出し動作の制御モードを設定するようにしてもよい。

【００２９】また、上述した実施の形態においては、キャッシュメモリ１１で保持し続ける必要がなくなったデータを通常の追い出し動作によって追い出すようにしているが、これに限らず、そのようなデータをキャッシュメモリ１１から直接削除したり、次回の追い出し動作時に強制的にキャッシュメモリ１１から追い出したりしてもよい。なお、後者の場合には、データの追い出し動作の制御モードの一つとして、キャッシュメモリ１１で保持される特定のデータを強制的に追い出すモードを新たに設け、このモードを管理テーブル１３の管理フラグ１３ａにて別途管理するようにするとよい。

【００３０】さらに、上述した実施の形態においては、キャッシュメモリ１１で新たに保持されるデータについて、その追い出し動作の制御モードを設定するようにしているが、これに限らず、キャッシュメモリ１１で既に保持されているデータについても、そのデータに対応する管理フラグ１３ａをＳＣＳＩメッセージ等によって変更するようにしてもよい。

【００３１】

【発明の効果】以上説明したように本発明によれば、キャッシュメモリで保持されるデータに対して外部から適切な指示を与えることにより、頻繁にアクセスされると予想されるデータを確実にキャッシュメモリに残すことができるので、キャッシュヒット率を向上させることができ、全体のデータ処理性能を高めることができる。

【図面の簡単な説明】

【図１】本発明によるキャッシュ記憶装置が適用される動画ストリームサーバの全体構成を示すブロック図。

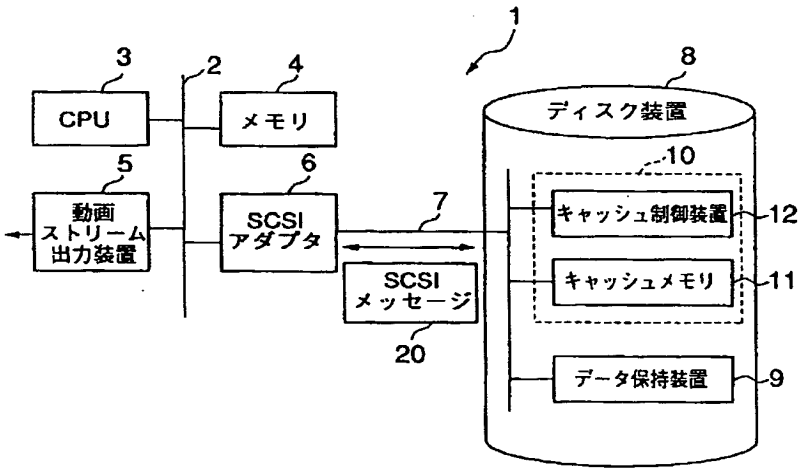
【図２】図１に示すキャッシュ記憶装置の論理構成を示す図。

【図３】キャッシュ記憶装置に対して与えられるＳＣＳＩメッセージの一例を示す図。

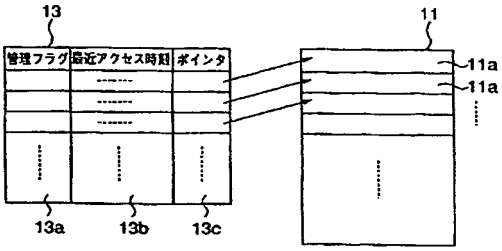
【符号の説明】

１　動画ストリームサーバ
２　メインバス
３　ＣＰＵ（中央処理装置）
４　メモリ
５　動画ストリーム出力装置
６　ＳＣＳＩアダプタ
７　ＳＣＳＩバス
８　ディスク装置
９　データ保持装置
１０　キャッシュ記憶装置
１１　キャッシュメモリ
１２　キャッシュ制御装置
１３　管理テーブル
１３ａ　管理フラグ
２０　ＳＣＳＩメッセージ

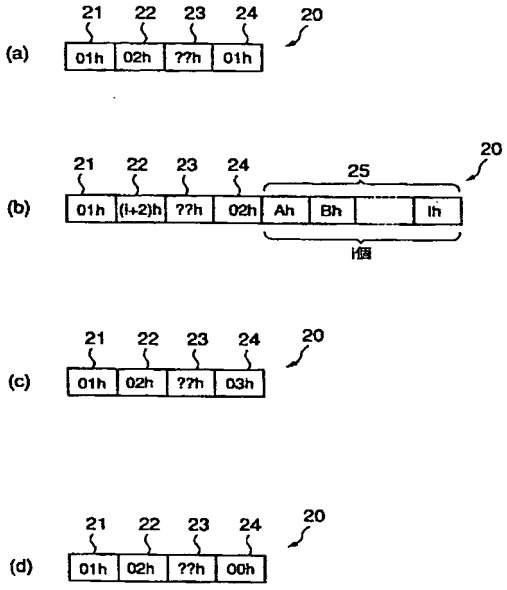
【図１】



【図２】



【図３】



フロントページの続き

(51) Int.Cl.[7]	識別記号	FI	テーマコード(参考)
G06F 12/08	320	G06F 12/08	320

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☑ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.